5. CONCLUSIONS AND RECOMMENDATIONS

In summary, the Data Inventory Project proved to be far more complicated and time-consuming than we originally envisioned. The original schedule was doubled by extensions, and it is apparent that work on the project will continue after the project's completion marked by this report. Several factors contributed to the time expansion:

(1) There proved to be a large number of data sources for Galveston Bay, but only a minority could be described as major projects (e.g., the TWC Statewide Monitoring Network, the Galveston Bay Project, the TWDB Bays & Estuaries Program, etc.), i.e. the data resource can be described as a few major projects and a great many small projects, which served to multiply contact time and logistics;

(2) The point-of-contact approach failed, requiring much greater time and effort of the PI's to find and gain access to agency data holdings;

(3) In general, the response of the data sources to our inquiries has been poor, requiring multiple letters or calls, and requiring months (at best) to finally gain access to data: many are only now responding.

However, the dominant reason is that the management of older data--and by this we mean any data taken prior to 1980--is by-and-large a shambles. Much more effort was needed to locate and retrieve this data than expected. For all of these reasons far more PI time has been dissipated by searching and agency communication, than has been invested in actual data inventory and data base creation. The frustration of the time-consumption of tracking down misplaced data sets has been compensated (somewhat) by the conviction that this work had to be done, and, as we began to realize as the project developed, the sooner the better.

Some definite conclusions regarding the data resource for Galveston Bay can be drawn. These conclusions apply primarily to data on the biological, water quality and hydrographic features of the system, which are the most important insofar as the GBNEP objectives are concerned.

1. Most of the data sets for Galveston Bay taken prior to 1980 are presently inaccessible. The majority of this data appears to be irrevocably lost .

2. When one considers that the data prior to 1980 comprises the vast majority of data taken in Galveston Bay ever, in terms of sampling intensity (though this is compensated somewhat by the greater number of observations per sample due to modern metrological and analytical technology), this implies that most of the data resource has vanished.

3. The factors which have led to this loss of data are still operating today.

These conclusions of course must be qualified for specificity. For example, sediment quality data is of more recent concern, and has benefited from advances in analytical technology, so is in relatively good shape. Also, specific data collections with national

archival procedures are well-managed, e.g. the historical mapping of the National Ocean Service and its predecessor agencies, and the data collection efforts of the U.S. Geological Survey. Further, several important data sets have been recovered since 31 December 1990, including the Galveston Bay Project High-Frequency Program and the USCE 1936-37 program, and some which were previously unavailable have now been provided. On the other hand, there are enough major data sets that remain lost, including the USBCF biological program of 1958-67, data collections by the Harris County Pollution Control Department from the 1960's through 1981, the intensive studies of the Houston Ship Channel by Texas A&M, and the intensive sampling performed by the City of Houston and by Harris County in the 1940's and earlier, that the above conclusions still hold. Further, these are examples of *agency* programs; the situation is worse for research data of individuals.

We have identified seven principal factors that contribute to this data loss, as follows:

1. Low priority assigned to archiving and preservation of older data.

This is a reflection of human psychology. Once a project or survey is completed, there is a tendency to stack the results out of the way and move on to the next challenge. Many agencies operate under a pressure of time, which conspires against good archival practices. Some agencies have some form of data management currently in place. While this is encouraging, it is also precarious, in that these programs are sensitive to shifts in organizational emphasis. An office purge is forever.

2. Mission-specific agency operation: perception of old data as "obsolete" and archiving as an unwarranted expense.

The Corps collects hydrographic or water quality data to support, e.g., navigation projects in place or in planning. Once a condition survey has been used to determine the need for dredging, once a decision on spoil disposal is made, once a project design is completed, the data sets employed in those activities are no longer needed. The mission of Texas Parks and Wildlife is to monitor the state of the coastal fisheries. The present condition is always primary. The Texas Water Commission and EPA are concerned with the present loadings of contaminants and the enforcement of water quality standards. The level of loadings a decade ago, or even last year, are rarely pertinent to that mission. And so it goes. The value of data diminishes quickly with age in these kinds of problem-specific operations. Yet it is these agencies that are largely responsible for the bulk of data collection within the Galveston Bay system.

3. Personnel turnover, combined with little or no documentation.

Only a handful of people in an agency generally has immediate familiarity with a data base. If the data base is not currently in use, this number will decline due to turnovers. When the last of these leave, the institutional memory goes with them. This was apparently the fate of the Galveston Bay Project data tape, described above, as well as numerous other programs in both the public and private sector. In some instances we had agencies deny that a sampling program ever took place (despite historical documentation to the contrary). This problem is most acutely manifested in the case of a single principal investigator at a university. Most of the

rare data sets we succeeded in locating for this project resulted from contacting (finally) the one or two persons remaining in the agency that knew something about the data. In one instance, the sole remaining contact died shortly after locating and transmitting the data to this project.

4. Agency instability, i.e. dissolution, merging, reorganization, displacement & relocation.

Some data sets have survived by dint of being undisturbed, until this Data Inventory Project located them. With an office move, as parcels, files and boxes are shifted about, the exposure to loss or discard is greatly increased. The disarray and haste usually typifying such moves contribute to a "clean-the-house" mentality, exacerbated by snap judgments on the part of personnel in no position to appraise the value of information. The decision is forced to consider data sets whose retention is already tenuous. Clearly, any sort of instability that leads to such shifting of materiel increases the probability of data loss. Aerial photography is particularly exposed to such loss because it has a monetary value as salvage, due to its silver content, which further conspires against its preservation.

5. Natural calamities (fires, floods, hurricanes) in poorly protected housing.

This problem speaks for itself. We have had a surprisingly large number of losses to such events, cf. Table 9. Ironically, it is the large, centralized, difficult-to-duplicate sets that are most exposed. The usual problems of water leakage, faulty wiring, and deterioration operate everywhere, but the Texas coastal zone--where most of the Galveston Bay data is housed--is exposed to extraordinary hazards. The human tendency is to disregard the risk of extreme hazard: we cannot help but note that the new Galveston Bay Information Center is located on Pelican Island.

6. Changes in data management technology, without upgrading of historical files.

This is a surprising factor, at least to these authors. There are several forms of this technological hazard. The first is simple technological obsolescence. At the time of data entry, punched cards and 8-track formats seemed to be fixed technology. Now, they are virtually unreadable. There is a transition period, of course, when newer technologies replace the old, but the task of upgrading formats of large, rarely used data files is onerous and of low priority. Then, with the same apparent suddenness of the demise of the LP and the Magcard, the technological hardware support is no longer available. At this writing, many data sets are being "stored" on floppy disks. In five years, they could be as unreadable as 8-inch floppies are today.

A second variety of this hazard is software obsolescence, in which the encoding is no longer readable. This ranges from discontinuation of a proprietary software, to loss of the description of coding formats. The prominent example of the former is System-2000 data bases. There are several examples of the latter, in which there exist tapes containing numerical data which can be read but whose meaning is no longer documented.

The third form of this hazard is due to the increasing information density of digital storage. As large data bases are compressed into smaller physical dimensions, the possibility of physical loss is increased: an errant electromagnetic field, small fire, or simple mislaying can wipe out the equivalent of reams of data. Probably the most prevalent form of this hazard is the acquisition of parity errors on an archival tape, and data garbling by stray magnetic fields. (Desk-top speakers seem to be an inviting flat surface upon which to stack floppies.) As new high-density media begin to appear, e.g. the compact disc, the possibility of simple physical loss becomes greater.

7. Proprietary attitude toward data by individual PI's.

This has been an endemic problem in academia, but it is also too frequently manifested in federal and state agencies. We will not propose to analyze the causes of this mentality, which may be rooted in the publish-or-perish environment, the paranoia of being "scooped" in some great insight gleaned from data analysis, the notion that "information is power," the view that one's data is valuable, and the view that one's data is worthless. We will observe, however, that many data sets exist in only their original form, in the possession of the person or agency which originally collected it, and are unavailable for the use of other investigators. This is a major source of the category of "inaccessible" in Table 9 and Figs. 7-9.

The important observation about all of these factors is that they are self-exacerbating and mutually reinforcing. Low priority of data management implies poor housing and careless data management practices, and increases the exposure to discard due to agency instability. The existence of only one or a few copies of a data set, and its possession by one or a few individuals increase the potential of loss due to natural or technological hazards. All of these factors are continuing to work at present, and are creating the potential for further loss of data, which will be lamented in the future. In our view, the problem is critical.

The facile--and fatuous--recommendation to correct the situation would be to eliminate the above seven factors. We would proffer the following specific recommendations, which we believe to be more pragmatic and to lie within the purview of the Galveston Bay National Estuary Program or its participating agencies.

1. All sponsored research projects (including consulting contracts and interagency contracts) should include a *requirement* for preparation of a data report documenting the *raw* measurements of the project. If a digitized version of the data base is part of the project, transmittal of a copy on an appropriate digital medium should also be required, with written (hard-copy) documentation of formats and software operation. Compliance with this requirement should be a condition for any future contracts. For public agencies, the data so transmitted should then be subjected to the requirement of public distribution given in (3) below.

2. All projects internal to an agency, performed by an agency staff, involving observations and measurements should require preparation of a data

report. If a digitized version of the data base is part of the project, a copy on the appropriate digital medium should also be required, with written (hard-copy) documentation of formats and software operation. For public agencies, the data so transmitted should then be subjected to the requirement of public distribution given in (3) below.

3. In public agencies, the release of the data report and digital copy should be made mandatory after a certain calendar period, e.g., six months. (If the data is still under review, it should be so marked, but being under review should not be used as a reason for delaying release.) Reimbursement for the expense of copying is appropriate, but the price should be reasonable. After all, the public has already paid for it once.

4. All agency files and materials should be marked with a destruction schedule by its originator. For measurements and raw data, at least, the files should be marked "permanent storage, not for destruction." In some agencies, smaller but equivalent words may be desirable.

5. At least one hard-copy record of every data set should be maintained. This might be raw data sheets, or might be a print-out of a digital data record. Also, even when a data set exists in a digitized data-management format (e.g., a data base management software form such as Lotus or dBase), a separate version in general encoding format (e.g., ASCII) should be maintained.

6. Data Inventory and Acquisition Projects should be sponsored as soon as practicable, either internal to an agency, or through external contract, to extend the present activity for Galveston Bay, and to secure similar data sets for the other Texas embayments and for the Texas coast. In particular, holdings in the following agencies and sites should be retrieved, organized and, where appropriate, digitized:

   the Texas Parks and Wildlife Olmeto warehouse

   the U.S. Corps of Engineers: Galveston District, the Texas area offices and the Waterways Experiment Station

   the National Marine Fisheries Service laboratories in Galveston

   the major research universities in the Texas coastal zone

   private engineering firms, surveying companies and aerial photographic services

   the U.S. Fish and Wildlife Service offices in Houston and Slidell

7. Some centralized, cooperative data storage and management facility is needed, one which is divorced from the separate mission-oriented state and federal agencies. Emphasis should be on competence of staffing and an

appropriate delineation of scope. The Texas Natural Resources Information System could become this entity, but it suffers from many problems, not the least of which is adequate and stable funding, which presently prevent its serving this function. This recommendation, of course, exceeds the jurisdiction of the GBNEP agencies, but could profit from the strong unanimous support of these agencies. It is, however, the only long-range solution that is evident to us.